A Comparative Study of Sentiment Analysis on Bangla Noisy Texts

Md. Rashedul Hasan Safa
rashed.safa57@gmail.com
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Ayesha Siddika
itsmesiddika@gmail.com
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Raihana Tabassum
tabassummeem24@gmail.com
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Shanuma Afrin Meghla
rahnuma.meghla@gmail.com
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Jayed Mohammad Barek
jayed.barek1995@gmail.com
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Mohammed Julfikar Ali Mahbub
mohammed.julfikar.ali.mahbub@g.bracu.ac.bd
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Md Motahar Mahtab
md.motahar.mahtab@g.bracu.ac.bd
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Annajiat Alim Rasel
annajiat@gmail.com
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

***Abstract*—The fact that people have sentiments is perhaps the most significant distinction between robots and humans. Researchers have been working on ways to imitate sentimentality in computers for decades. The majority of recent SA research in natural language processing (NLP) has concentrated on the English language. Because of the rich grammatical structure of the text, a few notable studies have been conducted in the Bangla language sector. It should also be highlighted that Bangla lacks a comprehensive dataset. As a consequence, current research projects including Bangla have failed to yield findings that are similar to those produced by researchers in other languages and reusable for future study. The motive of this paper is to find a model that can classify sentiments from the noisy Bangla dataset more effectively than the existing models. A recently introduced noisy Bangla text dataset has been inspected in our work. Three categorical machine learning models namely classical, neural network, and transformers that are prevalent in NLP tasks have been evaluated. The experimental outcome showed that the classical machine learning model Support Vector machine with merged n-gram feature extractors performed preferably in contrast to the models in the same category and other categories of approaches implemented.**

***Index Terms*—Sentiment Analysis, Bangla Natural Language Processing, Machine Learning, Noisy Bangla dataset**

I. Introduction

Sentiment Analysis is now the hottest, most demanding, and fastest-growing topic in the language processing industry. Sentiment analysis (SA) has acquired widespread study attention in natural language processing as a result of the growth of social digital material on the Internet. Sentiment analysis of phrases and sentences relies heavily on semantic construction. This involves determining the scope and effect of negation in reversing the polarity of sentiment, as well as measuring the impact of modifiers like degree adverbs and intensifiers in rescaling the emotion's strength. Sentiment analysis (SA), also known as opinion mining [1], is a branch of research that extracts people's feelings, attitudes, and emotions to forecast polarity in public opinion or textual data from microblogging sites [2] on a well-publicized issue. Researchers are rapidly getting interested in this issue as SA becomes a significant subject in natural language processing (NLP) in the machine learning sector, owing to the vast amount of opinionated data available on the Internet. People nowadays share their opinions on a given product or item via social media sites, newspapers, blogs, and other venues. In most SA, the algorithm calculated the number of good and negative emotion instances in a given

piece of text. Sentiment Analysis/Opinion Mining from natural language text is therefore an interdisciplinary and diverse AI topic. It aims to bridge the gap between highly emotional humans and sentimentally constrained computers by designing computational systems that can identify and react to human users' sentimental emotions.

Nearly 200 million people speak Bangla as their primary language, with 160 million of them being Bangladeshis [1]. Bangladeshi people are increasingly engaged in online activities such as connecting with friends and family via social media, expressing their opinions and thoughts on popular microblogging and social networking sites, sharing opinions and thoughts via comments on online news portals, and shopping online via online marketplaces and other similar applications. However, since such enterprises have little or no human-to-human connection, it is becoming more difficult for them to monitor and evaluate market trends, particularly when doing so by studying client reactions to their goods or services. Furthermore, sifting through each customer's comments and reviews to determine their thoughts is time-consuming and, in some circumstances, impossible, particularly given that, in this day and age of digital connection, a large amount of data is created extremely rapidly. As a result, automated sentiment analysis (SA) may play a critical role in improving efficiency and productivity in this situation. Opinion extraction, sentiment mining, opinion mining, subjectivity analysis, emotion analysis, review mining, and other names for SA are often used as machine learning applications in a variety of fields. The majority of research papers on SA are written in English, even though Bangla SA is still in its infancy. In comparison to a more conventional machine learning approaches, Deep Learning methods, particularly recurrent model-based deep learning models, have seen a lot of success in Natural Language Processing (NLP) in recent years [4].

In this paper, we present a comparative analysis among three different groups of machine learning models on the noisy Bangla text dataset. The noisy text includes words spelling distortion, ambiguity, and redundancy in text. We inspect the models to see how well the models can handle the analysis of a noisy text and classify sentiments of positive, negative, and neutral. Classical approaches, Neural Networks, and Transformers groups are executed to see which group of the model shows the preferable result. The paper is further structured as Section 2 includes former related works, Section 3 introduces the dataset, Section 4 specifies the methodology utilized, Section 5 explains the analysis of the result, and Section 6 wrap up the paper.

II. Related Work

Sentiment Analysis (SA) is an opinion mining research that examines people's attitudes, sentiments, assessments, and appraisals of social entities such as goods, services, people, organizations, and events. The Socio-eCommerce value can gain a lot through sentiment analysis. The researches in this field are expanding with momentous change. Numerous research has been established handling Bangla texts and syntaxes to develop benchmark in the field of Bangla NLP. This study [1] looked at six different types of emotions: joyful, sad, tender, thrilled, furious, and afraid. The performance of the Naive Bayes Classification Algorithm and the Topical Approach were compared, and the topical approach outperformed the other two approaches at both levels of scale.

Hoq et al. [3] identified emotion from Bangla texts, the authors created four models using a mix of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) and several Word Embeddings such as Embedding Layer, Word2Vec, Global Vectors (Glove), and Continuous Bag of Words (CBOW) (words, sentences). Happiness, anger, and sorrow are the three primary emotions that their models can identify. A study [4] was conducted on Bengali Sports news comments published in several newspapers to build a deep learning model that can classify a remark based on its mood. Khan et al. [5] proposed a method that classified Happy, Sad, Angry, Surprised, and Excited. Aside from these feelings, the study also covers two categories: Abusive and Religious. To train the dataset, a variety of machine learning approaches. Support Vector Machine (SVM) accuracy of 62 percent, Random Forest (RF) accuracy of 58 percent, K-Nearest Neighbors (KNN) accuracy of 55 percent, Nave Bayes (NB) accuracy of 52 percent, and Neural Network accuracy of 50 percent were employed. In this research [6], current state-of-the-art word embedding techniques Word2vec Skip-Gram and Continuous Bag of Words were tested for SA in Bangla, as well as a Word to Index model. The Word2vec Skip-Gram model beat other models, achieving an accuracy of 83.79 percent. Alam et al. [7] presented a paradigm for analyzing feelings in Bangla-language writings. They utilized Bangla comments to create a classification model in their proposal. Convolutional Neural Network is a kind of neural network that generates the model. The classifier model achieves a classification accuracy of 99.87 percent, which is 6.87 percent higher than the best Bangla sentiment classifier currently available. Alam et al. [8] reviewed of Bangla NLP or BNLP tasks, resources, and tools available to the research community was cited. The benchmark datasets were collected from various platforms for nine NLP tasks using current state-of-the-art algorithms (i.e., transformer-based models). Comparison of monolingual vs. multilingual models of various sizes to obtain comparative outcomes for the NLP tasks under consideration was shown. Sen et al. [9] proposed a method which Information Extraction, Machine Translation (MT), Named Entity Recognition (NER), Parsing, Parts of Speech (POS) Tagging, Question Answering System, Sentiment Analysis, Spam and Fake Detection, Text Summarization, Word Sense Disambiguation, and Speech Processing and Recognition are among the 11 areas they analyzed. Most research was found to be after 2015. Exploration of traditional, machine learning, and deep learning methodologies using various datasets, as well as the BNLP's limits, present, and future developments.

This paper [11] presented a ML based approach for extracting sentiment polarity (positive or negative) from Bengali book reviews. A corpus of 2000 evaluations on Bengali literature is

produced to examine the efficiency of the suggested approach. The unigram, bigram, and trigram characteristics were also taken into account in a comparative analysis using different algorithms (for instance logistic regression (LR), Naive Bayes, SVM, and SGD). The multinomial Naive Bayes procedure with unigram feature surpasses the other strategies with being 84-percent accurate on the test set, according to the results. With a focus on texts in Bangla, the authors examine numerous issues and solutions for sentiment analysis.

Detection of Bangla sentences that are multi-type abusive in nature was done in this work [12]. For the purpose of classification various machine learning as well as deep learning techniques were applied. It also introduced new rules of stemming that enhanced the result. [13] The author developed an affect list using WordNet from available word lists in English. The manual work of translation, disambiguation, and updating can be reduced using an automatic approach. This work creates a list of WordNet effects in Bangla which can be employed in processing related data. [14] N. Banik et.el proposed a comparative study using models that are supervised in nature (Naive Bayes, SVM, Logistic Regression, CNN, and LSTM) and detected toxic behaviors from a public Bangla dataset. It found supremacy of deep learning models where CNN showed 95.30% performance accuracy. [15] Data containing comments on Bangla, English and a mixture of Romanized Bangla from YouTube were used to classify emotions of positive, negative, and neutral. The work also includes detection of 6 variety of emotions from text The proper sentiment of Bangla sentence are generally understood from a complete sentence but the authors [19] skipped the use of sentences larger than thirty words. The emotions associated with a sentence can often be encountered at the end of the sentence or paragraph. Hence to understand the perfect sentiment sentence in length containing more than thirty words required to be taken into consideration. [16] Word2Vec was used to detect sentences or words of high rank. Proper vocabulary is needed for embedding layers. But in Bangla, there are insufficient datasets that are large and can help to create a huge stock of vocabulary. [17] Facebook user comments from groups of various domains were used to build a Bangla text corpus. The naive Bayes model was used by the author to classify the text emotions. [18] The author used CNN with an attention base. It was the first implementation of CNN along with attention for the purpose of SA that gained 72.06% accuracy. [19] The work achieved 94% accuracy when advanced layer LSTM was executed. [20] Deep learning approaches on two separate datasets were applied. The study portrayed a notable comparison between the executed methods. Sentiments or opinions of posts from the Bangla microblog were extracted and analyzed by Chowdhury et al. [21]. This study pointed out the text polarity as to whether the post was positive or negative. Support vector machine or SVM and Maximum entropy were used for experimenting on a variety of combinations of features. 93% accuracy was achieved by using SVM with unigrams and the emoticon feature. They pointed out that the feature emoticon could play a key role while using a binary classifier. A sizable dataset of Bangla text and also romanized Bangla text is provided in [22] by Hassan to conduct experiments of SA. The dataset contained a total of 9337 instances with 6698 normal Bangla text and 2639 romanized Bangla text. They implemented LSTM with binary and categorical cross-entropy as loss functions. 78% accuracy was achieved when using 2 categories to classify and 55 % accuracy was achieved when using 3 categories. Sentiment polarity recognition process that recognizes the polarity of a Bangla tweet using ML processes was presented by Sarkar et al. [23]. This approach was tested on a dataset of Bangla tweets, that was released for SAIL contest. The authors used a combination of features including the SentiWordnet and n-grams. They used a variety of combinations of features to train multinomial NB and SVM classifiers. The best performance was shown by the SVM with Sentiword and unigram features.

III. Dataset

The SentNoB dataset introduced in the paper [2] has been used. The dataset contains 15,728 instances that are public comments collected from different social media platforms covering 13 domains. The instances are annotated to one of the labels from positive, negative, and neutral. The Figure 1 shows the percentage distribution of positively , negatively and neutral instances in the dataset.

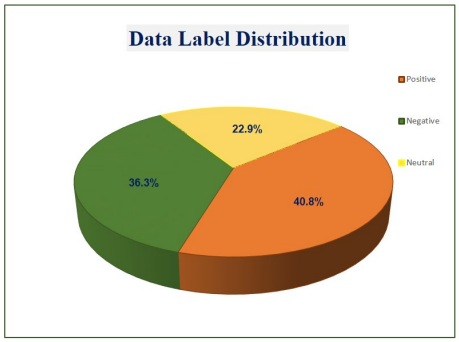

Fig. 1. Label (positive, negative and neutral) distribution in dataset.

A. Preprocessing

Preprocessing is a crucial step to go further in the modeling and analysis of data. Since the collected dataset has previously gone through manual noise and ambiguity reduction we have formally replaced the numerals with an SK token following [2]. The data set is split in the ratio of 80:20 where 80% is used for training, 10% for validation purposes, and 10% for testing.

IV. Experimentation

Several techniques have been implemented and analyzed for the purpose of text sentiment analysis. Starting from basic machine learning techniques to the pre-trained transformers Bangla texts have been assessed with the aim to automate the Bangla language tasks and tactics. Although, reaching

competitive benchmark on the data instances involving text from variety of areas is yet to achieve. In this work, we have applied three categories of machine learning models, Classical, Neural Networks, and Transformers to check the dominating category model on the SentNoB dataset. This work follows the way to Sentence level analysis to detect text sentiments expressing positive, negative, or neutral behavior of comments.

A. Classical Approach

The most basic machine learning algorithms were categorized in this group. The classical methods like Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) are employed. As the feature-based extractor combination of word n-grams (n=1 to 3) and character n-grams (n= 2 to 5) have shown improved F1-score in [2], we have integrated the feature-based extractor along with TF-IDF vectorizer to train the listed classical models.

B. Neural Network Techniques

Two of the highly utilized Recurrent Neural Network namely, Bi-LSTM and LSTM are applied from this category. We have followed the attention-based configuration to form the models as explained in [2] with random embedding and epoch size as 100 for both Bi-LSTM and LSTM.

C. Transformers

Transformer models are previously trained language methods that are fully based on the self-attention mechanism. Amongst the transformer models, multilingual models m-BERT and bn-BERT (bangla bert base) are selected and fine-tuned using the training split of the dataset over 100 epochs.The transformer models are bound to time complexities apart from having a unique pre-trained characteristics.

V. Analysis of Results

The three categorical implementations of algorithms are assessed by evaluation metrics precision, recall, and f1-score. Table I displays the respective metric scores of the models.

TABLE I
PERFORMANCE METRICS OF CLASSICAL, NEURAL NETWORK AND TRANSFORMER MODELS.

Methods		Precision	Recall	F1-score
Classical	Naive Bayes	51.38	65.57	57.62
	Random Forest	55.62	70.55	62.2
	Logistic Regression	57.05	73.2	64.13
	Support Vector Machine	57.9	73.27	64.68
Neural Network	LSTM	56.53	67.64	61.59
	Bi-LSTM	56.26	66.73	61.05
Transformers	m-BERT	46.99	56.68	51.39
	bn-BERT	48.19	58.76	52.95

An intuitive look reveals that the model performances are fluctuating between 50% to 65%. From the classical models, Support Vector Machine (SVM) has obtained high precision (57.9) and recall (73.27) resulting in a peak F1-score of 64.68. The classical approaches trained by vectors integrated with an n-gram feature extractor aids in the performance enhancement. In the neural network, LSTM showed top precision (56.53), recall (67.64), and f1-score (61.59) over Bi-LSTM with a slight difference. Amongst the transformer models, the bn-BERT acquired a high f1-score of 52.95. The f1-scores of all other models in sequence are Naive Bayes (57.62), Random Forest (62.2), Logistic Regression (64.13), and m-BERT (51.39). In contrast to the three categories, we found the transformers portrayed the lowest performance, the reason behind this can be the pre-trained corpus difference with that used in fine-tuning as also pointed in [2]. Hence through the analysis of the model outcome from each category, we can conclude that the classical SVM technique achieved superiority over all other models. The SVM integrated with feature extractors can handle the noisy Bangla dataset well in comparison to other groups of models.

VI. Conclusion

The rapid evolution of technology has led to much research on different behavioral aspects of humans from which natural language processing is one. Research in the language of different communities of the world is moving a step forward to development. Although it has been more than three decades of research on Bangla Natural Language Processing, the field is still to reach a satisfactory level due to complex syntactic analysis and lacking resources and tools. In this paper, we classified noisy Bangla text sentiments using a dataset containing comments and reviews from 13 separate areas. Three non-identical categories of machine learning techniques namely, Classical Machine Learning, Neural Network, and Transformer models are implemented. The classical approach SVM with a combination of feature extractors(word and character) gained supremacy over every other executed model. In the future, we will expand the dataset, inspect different preprocessing techniques and experiment with Bangla noisy dataset in large.

References

[1] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, A. K. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques", 2019 IEEE 4th International Conference on Computer and Communication Systems, pages 360-364, 2019.

[2] K. I. Islam, Md S. Islam, S. Kar, M. R. Amin, "SentNoB: A Dataset for Analysing Sentiment on Noisy Bangla Texts", November 2021.

[3] M. Hoq, P. Haque, M. N. Uddin, "Sentiment Analysis of Bangla Language Using Deep Learning Approaches", IEEE 15th International Conference on e-Business Engineering (ICEBE), pages 140-151, February 2021.

[4] M. Rahman, S. Haque, Z. R. Saurav, "Identifying and Categorizing Opinions Expressed in Bangla Sentences using Deep Learning Technique", International Journal of Computer Applications, vol- 176, pages 13-17, April 2020.

[5] Md. S. S. Khan, S. R. Rafa, Al E. H. Abir, A. K. Das, "Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity", Journal of Engineering Advancements, vol- 2, pages 118-124, July 2021.

[6] S. H. Sumit, Md. Z. Hossan, T. A. Muntasir, T. Sourov, “Exploring Word Embedding For Bangla Sentiment Analysis”, IInternational Conference on Bangla Speech and Language Processing(ICBSLP), pages 1-9,September 2018.

[7] Md. H. Alam, Md M. Rahoman, Md. A. K. Azad, “Sentiment Analysis for Bangla Sentences using Convolutional Neural Network”, 2017 20th International Conference of Computer and Information Technology (ICCIT), December 2017.

[8] FIROJ ALAM, MD. ARID HASAN, TANVIRUL ALAM, AKIB KHAN, JANNATUL TAJRIN, NAIRA KHAN, SHAMMUR ABSAR CHOWDHURY, “A Review of Bangla Natural Language Processing Tasks and the Utility of Transformer Models”, pages 1-48, July 2021.

[9] O. SEN, M. FUAD, MD. N. ISLAM, J. RABBI, MD.K. HASAN, M. BAZ, M. MASUD, MD. A. AWAL, A. A. FIME, , MD. T. H. FUAD, D. SIKDER, MD. A. RAIHAN, “Bangla Natural Language Processing: A Comprehensive Review of Classical, Machine Learning, and Deep Learning Based Methods”, vol- 4, 2017.

[10] E. Hossain, O. Sharif, M. M. Hoque, “Sentiment Polarity Detection on Bengali Book Reviews Using Multinomial Naïve Bayes”.

[11] M. A. Karim, M. Kaykobad, M. Murshed, “Technical Challenges and Design Issues in Bangla Language Processing”, pages 328-345.

[12] Emon, E.A., Rahman, S., Banarjee, J., Das, A.K., Mittra, T.: A deep learning approach to detect abusive Bengali text. In: 2019 7th International Conference on Smart Computing Communications (ICSCC),2019.

[13] Das, D., Bandyopadhyay, S.; ”Developing Bengali WordNet Affect for Analyzing Emotion ”,2010.

[14] Banik, N., Rahman, Md.;”Toxicity Detection on Bengali Social Media Comments using Supervised Models”; November 2019.

[15] Irtiza Tripto, N., Eunus Ali, M.; ”Detecting multilabel sentiment and emotions from Bangla youtube comments”; 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, pp. 1–6 ,2018.

[16] Abujar, S., Masum, A.K.M., Mohibullah, M., Ohidujjaman, Hossain, S.A.;” An approach for Bengali text summarization using Word2Vector.”; 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, pp. 1–5 ,2019.

[17] Azmin, S., Dhar, K.; ”Emotion detection from Bangla text corpus using Naïve Bayes classifier”.In: 2019 International Conference on Electrical Information and Communication Technology (ECIT), Khulna, 1–5 Dec 2019.

[18] Sharmin, S., Chakma, D.,”Attention-based convolutional neural network for Bangla sentiment analysis.”, AI Soc. 2020.

[19] Ahmed, A., Yousuf, M.A.,”Sentiment analysis on Bangla text using Long Short-Term Memory (LSTM) recurrent neural network.” Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Advances in Intelligent Systems and Computing, vol. 1309. Springer, Singapore ,2021.

[20] Rahman, M., Haque, S., Saurav, Z.R.,”Identifying and categorizing opinions expressed in Bangla sentences using deep learning technique.”, Int. J. Comput. Appl. 176(17), 0975–8887 ,2020.

[21] Chowdhury, Shaika and Chowdhury, Wasifa;”Performing sentiment analysis in Bangla microblog posts”;2014 International Conference on Informatics, Electronics & Vision (ICIEV), pages. 1-6,2014.

[22] Hassan, Asif and Amin, Mohammad Rashedul and Mohammed, N and Azad, AKA;”Sentiment analysis on bangla and romanized bangla text (BRBT) using deep recurrent models”,arXiv preprint arXiv:1610.00369, 2016.

[23] Sarkar, Kamal and Bhowmick, Mandira;”Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines”;2017 IEEE Calcutta Conference (CALCON),pages. 31-36,